

Midterm 2 — Sample 1

DATA 412/612

Practice exam, questions only. Cumulative through November 11 (weeks 1–11). The material since Midterm 1 is joins, strings, factors, dates, and statistics (a two-sample test and a linear model).

Put your answers in an R Markdown document and keep the code visible (`knitr::opts_chunk$set(echo = TRUE)` in the setup chunk). You may use your notes, the R help files, and the internet. Do not discuss the exam with anyone else. Load the tidyverse. `mpg` comes with `ggplot2`.

1. **Factors** (8 pts). In `mpg`, lump every `class` that appears fewer than 20 times into `"Other"`. Reorder the levels of that factor by the median `hwy`, lowest median first. Print the level order and a count of each level.
2. **Dates** (8 pts). Parse `stamp` as a date-time in the time zone `America/New_York`. Add the weekday name, the hour in that time zone, and the same instant displayed in UTC. Then add two days with a period, not a duration, and show the new calendar date.

```
stamps <- tibble(  
  stamp = c("2024-03-10 01:30", "2024-07-04 18:05")  
)
```

3. **A two-sample test** (8 pts). Using `mpg`, compare highway mileage for front-wheel drive (`drv == "f"`) with rear-wheel drive (`drv == "r"`). State the null hypothesis in one sentence. Report the difference in means and the p-value. Do not drop cars by hand before the test; let the test see those two groups.
4. **A linear model** (10 pts). Fit `hwy` on `displ` in `mpg`. In one sentence, say what the `displ` coefficient means. Plot residuals against fitted values and say whether the cloud looks like random scatter around zero.
5. **A join** (6 pts). Load `nycflights13`. Add the airline name onto `flights` and report the mean arrival delay for each airline name, ignoring missing delays. Show the three names with the largest mean delay.